Information Theoretic Model Selection for Pattern Analysis

Joachim M. Buhmann JBUHMANN@INF.ETHZ.CH
Morteza Haghir Chehreghani MORTEZA.CHEHREGHANI@INF.ETHZ.CH
Mario Frank MARIO.FRANK@INF.ETHZ.CH
Andreas P. Streich ANDREAS.STREICH@ALUMNI.ETHZ.CH
Department of Computer Science, ETH Zurich, Switzerland

Editor: I. Guyon, G. Dror, V. Lemaire, G. Taylor, and D. Silver

Abstract

Exploratory data analysis requires (i) to define a set of patterns hypothesized to exist in the data, (ii) to specify a suitable quantification principle or cost function to rank these patterns and (iii) to validate the inferred patterns. For data clustering, the patterns are object partitionings into k groups; for PCA or truncated SVD, the patterns are orthogonal transformations with projections to a low-dimensional space. We propose an information theoretic principle for model selection and model-order selection. Our principle ranks competing pattern cost functions according to their ability to extract context sensitive information from noisy data with respect to the chosen hypothesis class. Sets of approximative solutions serve as a basis for a communication protocol. Analogous to Buhmann (2010), inferred models maximize the so-called approximation capacity that is the mutual information between coarsened training data patterns and coarsened test data patterns. We demonstrate how to apply our validation framework by the well-known Gaussian mixture model and by a multi-label clustering approach for role mining in binary user privilege assignments.

Keywords: Unsupervised learning, data clustering, model selection, information theory, maximum entropy, approximation capacity

1. Model Selection via Coding

Model selection and model order selection [Burnham and Anderson (2002)] are fundamental problems in pattern analysis. A variety of models and algorithms has been proposed to extract patterns from data, i.e., for clustering, but a comprehensive theory on how to choose the "right" pattern model given the data is still missing. Statistical learning theory as in Vapnik (1998) advocates to measure the generalization ability of models and to employ the prediction error as a measure of model quality. In particular, stability analysis of clustering solutions has shown very promising results for model order selection in clustering [Dudoit and Fridlyand (2002); Lange et al. (2004)], although discussed controversially by Ben-David et al. (2006). Stability is, however, only one aspect of statistical modeling, e.g., for unsupervised learning. The other aspect of the modeling tradeoff is characterized by the informativeness of the extracted patterns. A tolerable decrease in the stability of inferred patterns in the data (data model) might be compensated by a substantial increase of their information content (see also the discussion in Tishby et al. (1999)).

We formulate a principle that balances these two antagonistic objectives by transforming the model selection problem into a coding problem for a generic communication scenario. Thereby, the objective function or cost function that maps patterns to quality scores is considered as a noisy channel. Different objectives are ranked according to their transmission properties and the cost function with the highest channel capacity is then selected as the most informative model for a given data set. Thereby, we generalize the set-based coding scheme proposed by Buhmann (2010) to sets of weighted hypotheses in order to simplify the embedding of pattern inference problems in a communication framework. Learning, in general, resembles communication from a conceptual viewpoint: For *communication*, one demands a high rate (a large amount of information transferred per channel use) together with a decoding rule that is stable under the perturbations of the messages by the noise in the channel. For *learning* patterns in data, the data analyst favors a rich model with high complexity (e.g., a large number of clusters in grouping), while the generalization error of test patterns is expected to remain stable and low. We require that solutions of pattern analysis problems are reliably inferred from noisy data.

This article first summarizes the information theoretic framework of weighted approximation set coding (wASC) for validating statistical models. We then demonstrate, for the first time, how to practically transform a pattern recognition task into a communication setting and how to compute the capacity of clustering solutions. The feasibility of wASC is demonstrated for mixture models and real world data.

2. Brief Introduction to Approximation Set Coding

In this section, we briefly describe the theory of weighted Approximation Set Coding (wASC) for pattern analysis as proposed by Buhmann (2010).

Let $\mathbf{X} = \{X_1, \ldots, X_n\} \in \mathcal{X}$ be a set of n objects $\mathbf{O}$ and n measurements in a data space $\mathcal{X}$, where the measurements characterize the objects. Throughout the paper, we assume the special case of a bijective map between objects and measurements, i.e., the i^{th} object is isomorphic to the vector $\mathbf{x}_i \in \mathbb{R}^D$. In general, the (object, measurement) relation might be more complex than an object-specific feature vector.

A **hypothesis**, i.e. a solution of a pattern analysis problem, is a function c that assigns objects (e.g. data) to patterns of a pattern space $\mathcal{P}$:

$$c \;:\; \mathcal{X} \to \mathcal{P}, \qquad \mathbf{X} \mapsto c(\mathbf{X}). \tag{1}$$

Accordingly, the **hypothesis class** is the set of all such functions, i.e. $\mathcal{C}(\mathbf{X}) := \{c(\mathbf{X}) : \mathbf{X} \in \mathcal{X}\}$. For clustering, the patterns are object partitionings $\mathcal{P} = \{1, \ldots, k\}^n$. A model for pattern analysis is characterized by a **cost** or **objective function** $R(c, \mathbf{X})$ that assigns a real value to a pattern $c(\mathbf{X})$. To simplify the notation, model parameters $\boldsymbol{\theta}$ (e.g., centroids) are not explicitly listed as arguments of the objective function. Let $c^\perp(\mathbf{X})$ be the pattern that minimizes the cost function, i.e. $c^\perp(\mathbf{X}) \in \arg\min_c R(c, \mathbf{X})$. As the measurements $\mathbf{X}$ are random variables, the global minimum $c^\perp(\mathbf{X})$ of the empirical costs is a random variable as well. In order to rank all solutions of the pattern analysis problem, we introduce *approximation weights*

$$w : \mathcal{C} \times \mathcal{X} \times \mathbb{R}_+ \to [0,1]\,, \qquad (c, \mathbf{X}, \beta) \mapsto w_\beta(c, \mathbf{X})\,. \tag{2}$$

The weights are chosen to be non-negative $w_\beta(c, \mathbf{X}) \geq 0$, the maximal weight is allocated to the global minimizer $c^\perp$ and it is normalized to one ($w_\beta(c^\perp, \mathbf{X}) = 1$). Semantically, the weights $w_\beta(c, \mathbf{X})$ quantify the quality of a solution w.r.t. the global minimizer of $R(., \mathbf{X})$. The scaling parameter β controls the size of the solution set. Large β yields a small solution set and small β renders many solutions as good approximations of the minimizer $c^\perp(\mathbf{X})$ in terms of costs, i.e., $w_\beta(c, \mathbf{X}) > 1 - \epsilon$ denotes that c is regarded as an ϵ/β-good approximation of the minimal costs $R(c^\perp, \mathbf{X})$. Therefore, we also require that weights fulfil the inverse order constraints compared to costs, i.e., a solution c with lower or equal costs than $\tilde{c}$ should have a larger or equal weight, i.e.,

$$R(c, \mathbf{X}) \leq R(\tilde{c}, \mathbf{X}) \quad \Longleftrightarrow \quad w_\beta(c, \mathbf{X}) \geq w_\beta(\tilde{c}, \mathbf{X}) \; . \tag{3}$$

Given a cost function $R(c, \mathbf{X}))$ these order constraints determine the weights up to a monotonic (possibly nonlinear) transformation $f(.)$ which effectively rescales the costs ($\tilde{R}(c, \mathbf{X}) = f(R(c, \mathbf{X})))$. The family of (Boltzmann) weights

$$w_\beta(c, \mathbf{X}) := \exp\bigl(-\beta \Delta R(c, \mathbf{X})\bigr) \; , \text{ with } \Delta R(c, \mathbf{X}) := R(c, \mathbf{X}) - R(c^\perp, \mathbf{X})) \tag{4}$$

parameterized by the inverse computational temperature β, fulfils these requirements. Although the Boltzmann weights are a special choice, all other weighting schemes can be explained by a monotonic rescaling of the costs.

Conceptually, wASC assumes a “two sample set scenario” as in Tishby et al. (1999). Let $\mathbf{X}^{(q)}, q \in \{1, 2\}$, be two datasets with the same inherent structure but different noise instances. In most cases, their sets of global minima differ, i.e. $\{c^\perp(\mathbf{X}^{(1)})\} \cap \{c^\perp(\mathbf{X}^{(2)})\} = \emptyset$, demonstrating that the global minimizers often lack robustness to fluctuations. The approximation weights (2) have been introduced to cure this instability. Solutions with large approximation weights $w_\beta(c, \mathbf{X}) \geq 1 - \epsilon, \; \epsilon \ll 1$ can be accepted as substitutes of the global minimizers. Adopting a learning theoretic viewpoint, the set of solutions with large weights generalizes significantly better than the set of global minimizers, provided that β is suitably chosen. The concept wASC serves the purpose to determine such an appropriate scale β. The two data sets $\mathbf{X}^{(q)}, q \in \{1, 2\}$, define two weight sets $w_\beta(c, \mathbf{X}^{(q)})$. These weights give rise to the two *weight sums* $\mathcal{Z}_q$ and the *joint weight sum* $\mathcal{Z}_{12}$

$$\begin{aligned} \mathcal{Z}_q := \mathcal{Z}(\mathbf{X}^{(q)}) \quad &= \sum_{c \in \mathcal{C}(\mathbf{X}^{(q)})} \exp\bigl(-\beta \Delta R(c, \mathbf{X}^{(q)})\bigr), \; q = 1, 2 && (5) \\ \mathcal{Z}_{12} := \mathcal{Z}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) \quad &= \sum_{c \in \mathcal{C}(\mathbf{X}^{(2)})} \exp\bigl(-\beta(\Delta R(c, \mathbf{X}^{(1)}) + \Delta R(c, \mathbf{X}^{(2)}))\bigr), && (6) \end{aligned}$$

where $\exp(-\beta(\Delta R(c, \mathbf{X}^{(1)}) + \Delta R(c, \mathbf{X}^{(2)})))$ measures how well a solution c minimizes costs on *both* datasets. The sums (5,6) play a central role in our framework. If $\beta = 0$, all weights $w_\beta(c, \mathbf{X}) = 1$ are independent of the costs. In this case, $\mathcal{Z}_q = |\mathcal{C}(\mathbf{X}^{(q)})|$ indicates the size of the hypothesis space, and $\mathcal{Z}_{12} = \mathcal{Z}_1 = \mathcal{Z}_2$. For high β, all weights are small compared to the weight $w_\beta(c^\perp, \mathbf{X}^{(q)})$ of the global optimum and the weight sum essentially counts the number of globally optimal solutions. For intermediate β, $\mathcal{Z}(\cdot)$ takes a value between 0 and $|\mathcal{C}(\mathbf{X}^{(q)})|$, giving rise to the interpretation of $\mathcal{Z}(\cdot)$ as the effective number of patterns that approximately fit the dataset $\mathbf{X}^{(q)}$, where β defines the precision of this approximation.

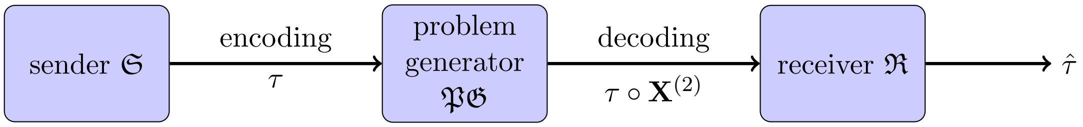


Figure 1: Communication process: (1) the sender selects transformation τ, (2) the problem generator draws $\mathbf{X}^{(2)} \sim \mathbb{P}(\mathbf{X})$ and applies τ to it, and (3) the receiver estimates $\hat{\tau}$ based on $\tilde{\mathbf{X}} = \tau \circ \mathbf{X}^{(2)}$.

Essentially, $\mathcal{Z}_q$ counts all statistically indistinguishable data patterns that approximate the minimum of the objective function. The global optimum $c^{\perp}(\mathbf{X})$ can change whenever we optimize on another random subset of the data, whereas, for a well-tuned β, the set of weights $\{w_\beta(c, \mathbf{X})\}$ remains approximately invariant. Therefore, β defines the resolution of the hypothesis class that is relevant for inference. Noise in the measurements $\mathbf{X}$ reduces this resolution and thus coarsens the hypothesis class. As a consequence, the key problem of learning is to control the resolution optimally: How high can β be chosen to still ensure identifiability of $\{w_\beta : w_\beta(c, \mathbf{X}) \geq 1 - \epsilon\}$ in the presence of data fluctuations? Conversely, choosing β too low yields a too coarse resolution of solutions and does not capture the maximal amount of information in the data.

We answer this key question by means of a communication scenario. The communication architecture includes a *sender* $\mathfrak{S}$ and a *receiver* $\mathfrak{R}$ with a *problem generator* $\mathfrak{PG}$ between the two terminals $\mathfrak{S}$, $\mathfrak{R}$ (see Fig. 1). The communication protocol is organized in two stages: (i) design of a communication code and (ii) the communication process.

For the **communication code**, we adapt Shannon's random coding scenario, where a codebook of random bit strings covers the space of all bit strings. In random coding, the sender sends a bit string and the receiver observes a perturbed version of this bit string. For decoding, the receiver has to find the most similar codebook vector in the codebook which is the decoded message. In the same spirit, for our scenario, the sender must communicate patterns to the receiver via noisy datasets. Since we are interested in patterns with low costs, the optimal pattern $c^{\perp}(\mathbf{X}^{(1)})$ can serve as a message. The other patterns in the codebook are generated by transforming the training data $\tau \circ \mathbf{X}^{(1)}$ with the transformation $\tau \in \mathbb{T} := \{\tau_1, ..., \tau_{2^{n\rho}}\}$. The number of codewords is $2^{n\rho}$ and ρ is the rate of the protocol. The choice of such transformations depends on the hypothesis class and they have to be equivariant, i.e., the transformed optimal pattern equals the optimal pattern of the transformed data $\tau \circ c(\mathbf{X}^{(1)}) = c(\tau \circ \mathbf{X}^{(1)})$. In data clustering, *permuting* the indices of the objects defines the group of transformations to cover the pattern space. Each clustering solution $c \in \mathcal{C}(\mathbf{X}^{(1)})$ can be transformed into another solution by a permutation τ on the indices of c.

To **communicate**, $\mathfrak{S}$ selects a transformation $\tau_s \in \mathbb{T}$ and sends it to a *problem generator* $\mathfrak{PG}$ as depicted in Fig. 1. $\mathfrak{PG}$ then generates a new dataset $\mathbf{X}^{(2)}$, applies the transformation τ_s, and sends the resulting data $\tilde{\mathbf{X}} := \tau_s \circ \mathbf{X}^{(2)}$ to $\mathfrak{R}$. On the receiver side, the lack of knowledge on the transformation τ_s is mixed with the stochastic variability of the source generating the data $\mathbf{X}$. $\mathfrak{R}$ has to estimate the transformation $\hat{\tau}$ based on $\tilde{\mathbf{X}}$. The decoding rule of $\mathfrak{R}$ selects the pattern transformation $\hat{\tau}$ that yields the highest joint weight sum of

$\hat{\tau} \circ \mathbf{X}^{(1)}$ and $\tilde{\mathbf{X}}$, i.e.,

$$\hat{\tau} \in \arg\max_{\tau \in \mathbb{T}} \sum_{c \in \mathcal{C}(\mathbf{X}^{(1)})} \exp(-\beta(R(c, \tau \circ \mathbf{X}^{(1)}) + R(c, \tilde{\mathbf{X}}))) \ . \tag{7}$$

In the absence of noise in the data, we have $\mathbf{X}^{(1)} = \mathbf{X}^{(2)}$, and error-free communication works even for $\beta \to \infty$. The higher the noise level, the lower we have to choose β in order to obtain weight sums that are approximately invariant under the stochastic fluctuations in the measurements thus preventing decoding errors. The error analysis of this protocol investigates the probability of decoding error $\mathbb{P}(\hat{\tau} \neq \tau_s | \tau_s)$. As derived for an equivalent channel in Buhmann (2011), an asymptotically vanishing error rate is achievable for rates

$$\rho \ \leq \ \mathcal{I}_\beta(\tau_s, \hat{\tau}) \ = \ \frac{1}{n} \log\left(\frac{|\{\tau_s\}| \mathcal{Z}_{12}}{\mathcal{Z}_1 \cdot \mathcal{Z}_2}\right) \ = \ \frac{1}{n}\left(\log \frac{|\{\tau_s\}|}{\mathcal{Z}_1} + \log \frac{|\mathcal{C}^{(2)}|}{\mathcal{Z}_2} - \log \frac{|\mathcal{C}^{(2)}|}{\mathcal{Z}_{12}}\right) \tag{8}$$

The three logarithmic terms in eq.(8) denote the mutual information between the coarsening of the pattern space on the sender side and the coarsening of the pattern space on the receiver side.

The cardinality $|\{\tau_s\}|$ is determined by the number of realizations of the random transformation τ, i.e. by the entropy of the type (in an information theoretic sense) of the empirical minimizer $c^\perp(\mathbf{X})$. As the entropy increases for a large number of patterns, $|\{\tau_s\}|$ accounts for the model complexity or informativeness of the solutions. For noisy data, the communication rate is reduced as otherwise the solutions can not be resolved by the receiver. The relative weights are determined by the term $\mathcal{Z}_{12}/(\mathcal{Z}_1 \cdot \mathcal{Z}_2) \in [0, 1]$ which accounts for the stability of the model under noise fluctuations.

In analogy to information theory, we define the *approximation capacity* as

$$\mathcal{CAP}(\tau_s, \hat{\tau}) = \max_\beta \mathcal{I}_\beta(\tau_s, \hat{\tau}) \, . \tag{9}$$

Using these entities, we can describe how to apply the wASC principle for model selection from a set of cost functions $\mathcal{R}$: Randomly split the given dataset $\mathbf{X}$ into two subsets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. For each candidate cost function $R(c, \mathbf{X}) \in \mathcal{R}$, compute the mutual information (eq. 8) and maximize it with respect to β. Then select the cost function that achieves highest capacity at the best resolution $\beta^\star$.

There exists a long history of information theoretic approaches to model selection, which traces back at least to Akaike's extension of the Maximum Likelihood principle. AIC penalizes fitted models by twice the number of free parameters. The Bayesian Information Criterion (BIC) suggests a stronger penalty than AIC, i.e., number of model parameters times logarithm of the number of samples. Rissanen's minimum description length principles is closely related to BIC (see e.g. Hastie et al. (2008) for model selection penalties). Tishby et al. (1999) proposed to select the number of clusters according to a difference of mutual informations which they called the imformation bottleneck. This asymptotic concept is closely related to rate distortion theory with side information (see Cover and Thomas (2006)). Finite sample size corrections of the information bottleneck allowed Still and Bialek (2004) to determine an optimal temperature with a prefered number of clusters.

3. Approximation Capacity for Parametric Clustering Models

Let $\mathbf{X}^{(q)}, q \in \{1, 2\}$ be two datasets drawn from the same source. We consider a parametric clustering model with K clusters. Then the cost function can be written as

$$R(c, \mathbf{X}) = \sum_{i=1}^{n} \epsilon_{i,c(i)} \text{ with } \forall i,\ c(i) \in \{1, .., K\}\,. \tag{10}$$

$\epsilon_{i,c(i)}$ indicates the costs of assigning object i to cluster $c(i)$. These costs $\epsilon_{i,c(i)}$ also contains all relevant parameters to identify a clustering solution, e.g. centroids. In the well-known case of k-means clustering we derive $\epsilon_{i,c(i)} = \|x_i - y_{c(i)}\|^2$.

Calculating the approximation capacity requires the following steps:

1. Identify the hypothesis space of the models and compute the cardinality of the set of possible transformations $|\{\tau_s\}|$.
2. Calculate the weight sums $\mathcal{Z}_q\,,\ q = 1, 2$, and the joint weight sum $\mathcal{Z}_{12}$.
3. Maximize $\mathcal{I}_\beta$ in Eq. (8) with respect to β.

In clustering problems, the hypothesis space is spanned by all possible assignments of objects to sources. The appropriate transformation in clustering problems is the permutation of objects. Albeit a solution contains the cluster assignments *and* cluster parameters like centroids, the centroid parameters contribute almost no entropy to the solution. With given cluster assignments the solution is fully determined as the objects of each cluster pinpoint the centroids to a particular vector. With the permutation transformations one can construct all clusterings starting from a single clustering. However, as the mutual information in Eq. (8) is estimated solely based on the identity transformation, one can ignore the specific kind of transformations when computing this estimate. The cardinality $|\{\tau_s\}|$ is then the number of all distinct clusterings on $\mathbf{X}^{(1)}$.

We obtain the individual weight sums and the joint weight sum by summing over all possible clustering solutions

$$\mathcal{Z}_q \quad = \quad \sum_{c \in C(\mathbf{X}^{(q)})} \exp\left(-\beta \sum_{i=1}^{n} \epsilon^{(q)}_{i,c(i)}\right) = \prod_{i=1}^{n} \sum_{k=1}^{K} \exp\left(-\beta \epsilon^{(q)}_{i,k}\right)\,, q = 1, 2, \tag{11}$$

$$\mathcal{Z}_{12} \quad = \quad \sum_{c \in C(\mathbf{X}^{(2)})} \exp\left(-\beta \sum_{i=1}^{n} (\epsilon^{(1)}_{i,c(i)} + \epsilon^{(2)}_{i,c(i)})\right) = \prod_{i=1}^{n} \sum_{k=1}^{K} \exp\left(-\beta \big(\epsilon^{(1)}_{i,k} + \epsilon^{(2)}_{i,k}\big)\right). \tag{12}$$

By substituting these weight sums to Eq. (8), the mutual information amounts to

$$\mathcal{I}_\beta = \frac{1}{n} \log |\{\tau_s\}| + \frac{1}{n} \sum_{i=1}^{n} \left(\log \sum_{k=1}^{K} e^{-\beta\left(\epsilon^{(1)}_{i,k} + \epsilon^{(2)}_{i,k}\right)} - \log \sum_{k=1}^{K} e^{-\beta \epsilon^{(1)}_{i,k}} \sum_{k'=1}^{K} e^{-\beta \epsilon^{(2)}_{i,k'}} \right). \tag{13}$$

The approximation capacity is numerically determined as the maximum of $\mathcal{I}_\beta$ over β.

4. Approximation Capacity for Mixtures of Gaussians

In this section, we demonstrate the principle of maximum approximation capacity on the well known Gaussian mixture model (GMM). We first study the approximation set coding for GMMs and then we experimentally compare it against other model selection principles.

4.1. Experimental Evaluation of Approximation Capacity

A GMM with K components is defined as $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \; \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \mathbf{\Sigma})$, with non-negative π_k and $\sum_k \pi_k = 1$. For didactical reasons, we do not optimize the covariance matrix $\mathbf{\Sigma}$ and simply fix it to $\mathbf{\Sigma} = 0.5 \cdot \mathbf{I}$. Then, maximizing the GMM likelihood essentially reduces to centroid-based clustering. Therefore, $\epsilon_{i,k} := \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$ indicates the costs of assigning object i to cluster k.

For experimental evaluation, we define $K = 5$ Gaussians with parameters $\pi_k = 1/K$, $\boldsymbol{\mu} \in \{(1, 0), (0, 1.5), (-2, 0), (0, -3), (4.25, -4)\}$, and with covariance $\mathbf{\Sigma} = 0.5 \cdot \mathbf{I}$. Let $\mathbf{X}^{(q)}, q \in \{1, 2\}$ be two datasets of identical size $n = 10,000$ drawn from these Gaussians. We optimize the assignment variables and the centroid parameters of our GMM model via annealed Gibbs sampling [Geman and Geman (1984)]. The computational temperature in Gibbs sampling is equivalent to the assumed width of the distributions. Thereby, we provide twice as many clusters to the model in order to enable overfitting. Starting from a high temperature, we successively cool down while optimizing the model parameters. In Figure $2(a)$, we illustrate the positions of the centroids with respect to the center of mass. At high temperature, all centroids coincide, indicating that the optimizer favors one cluster. As the temperature is lowered further, the centroids separate into increasingly many clusters until, finally, the sampler uses all 10 clusters to fit the data.

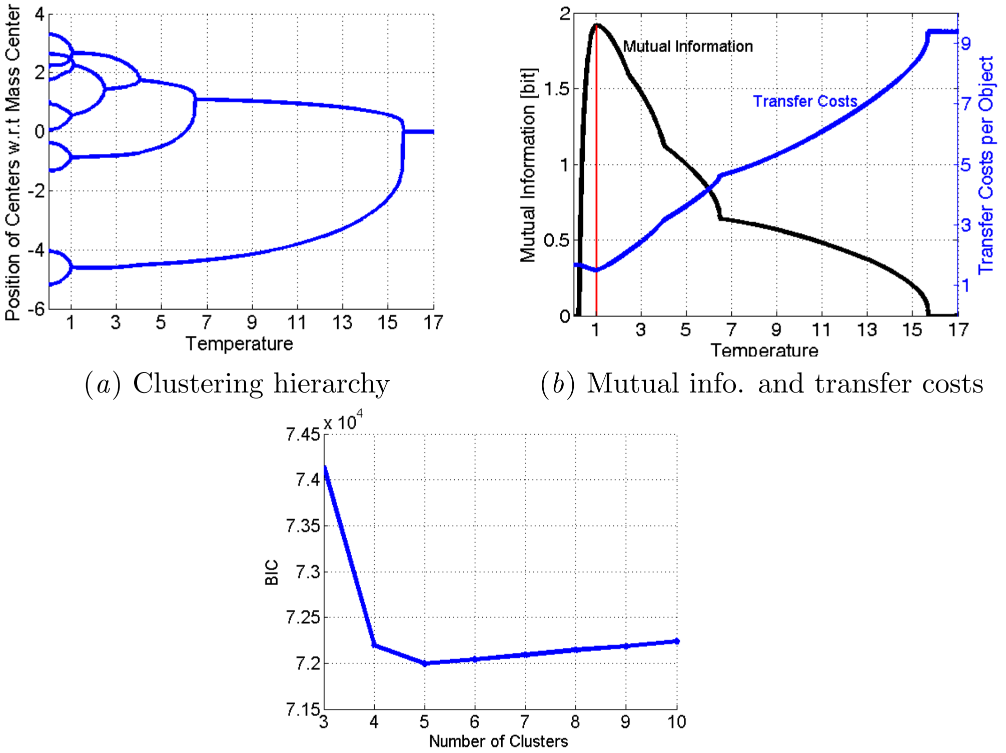
Figure $2(b)$ shows the numerical analysis of the mutual information in Eq. (13). When the stopping temperature of the Gibbs sampler coincides with the temperature β^{-1} that maximizes mutual information, we expect the best tradeoff between robustness and informativeness. And indeed, as illustrated in Figure $2(a)$, the correct model-order $\hat{K} = 5$ is found at this temperature. At lower stopping temperatures, the clusters split into many instable clusters which increases the decoding error, while at higher temperatures informativeness of the clustering solutions decreases.

4.2. Comparison with other principles

We compare approximation capacity against two other model order selection principles: i) generalization ability, and ii) BIC score.

Relation to generalization ability: A properly regularized clustering model explains not only the dataset at hand, but also new datasets from the same source. The inferred model parameters and assignment probabilities from the first dataset $\mathbf{X}^{(1)}$ can be used to compute the costs for the second dataset $\mathbf{X}^{(2)}$. The appropriate clustering model yields low costs on $\mathbf{X}^{(2)}$, while very informative but unstable structures and also very stable but little informative structures have high costs due to overfitting or underfitting, respectively.

We measure this generalization ability by computing the "transfer costs" $R(c^{(1)}, \mathbf{X}^{(2)})$ [Frank et al. (2011)]: At each stopping temperature of the Gibbs sampler, the current



(a) Clustering hierarchy

(b) Mutual info. and transfer costs

(c) BIC measure

Figure 2: Annealed Gibbs sampling for GMM: Influence of the stopping temperature for annealed optimization on the mutual information, on the transfer costs and on the positions of the cluster centroids. The lowest transfer cost is achieved at the temperature with highest mutual information. This is the lowest temperature at which the correct number of clusters $\hat{K} = 5$ is found. The hierarchy in Fig. 2(a) is obtained by projecting the two-dimensional centroids at each stopping temperature to the optimal one-dimensional subspace using multidimensional scaling. BIC verifies correctness of $\hat{K} = 5$.

parameters $\boldsymbol{\mu}^{(1)}$ and assignment probabilities $\mathbf{P}^{(1)}$ inferred from $\mathbf{X}^{(1)}$ are transfered to $\mathbf{X}^{(2)}$. The assignment probabilities $\mathbf{P}^{(1)}$ assume the form of a Gibbs distribution

$$p(\boldsymbol{\mu}_k^{(1)}|\mathbf{x}_i^{(1)}) = Z_x^{-1} \exp\left(-\beta\|\mathbf{x}_i^{(1)} - \boldsymbol{\mu}_k^{(1)}\|^2\right), \tag{14}$$

with Z_x as the normalization constant. The expected transfer costs with respect to these probabilities are then

$$\left\langle R(c^{(1)}, \mathbf{X}^{(2)})\right\rangle = \sum_{i=1}^{n}\sum_{k=1}^{K} p(\mathbf{x}_i^{(1)}, \boldsymbol{\mu}_k^{(1)})\|\mathbf{x}_i^{(2)} - \boldsymbol{\mu}_k^{(1)}\|^2 \approx \frac{1}{n}\sum_{n=1}^{n}\sum_{k=1}^{K} p(\boldsymbol{\mu}_k^{(1)}|\mathbf{x}_i^{(1)})\|\mathbf{x}_i^{(2)} - \boldsymbol{\mu}_k^{(1)}\|^2 , \tag{15}$$

Figure 2(b) illustrates the transfer costs as a function of β and compares it with the approximation capacity. The optimal transfer costs are obtained at the stopping temperature that corresponds to the approximation capacity.

Relation to BIC Arguably the most popular criterion for model-order selection is BIC as proposed by Schwarz (1978). It is, like wASC, an asymptotic principle, i.e. for sufficiently many observations, the fitted model preferred by BIC ideally corresponds to the candidate which is a posteriori most probable. However, the application of BIC is limited to models where one can determine the number of free parameters as here with GMM. Figure 2(c) confirms the consistency of wASC with BIC in finding the correct model order in our experiment.

5. Model Selection for Boolean matrix factorization

To proceed with studying different applicability aspects of wASC, we now consider the task to select one out of four models for factorizing a Boolean matrix with the clustering method proposed in Streich et al. (2009). Experiments with known ground truth allow us to rank these models according to their parameter estimation accuracy. We investigate whether wASC reproduces this ranking.

5.1. Data and Models

Consider binary data $\mathbf{X} \in \{0,1\}^{n \times D}$ in D dimensions, where a row $\mathbf{x}_i$ describes a single data item. Each data item i is assigned to a set of sources $\mathcal{L}_i$, and these sources generate the measurements $\mathbf{x}_i$ of the data item. The probabilities $v_{k,d}$ of a source k to emit a zero in dimension d parameterize the sources. To generate a data item i, one sample is drawn from each source in $\mathcal{L}_i$. In each dimension d, the individual samples are then combined via the Boolean OR to obtain the structure part of the data item $\tilde{\mathbf{x}}_i$. Finally, a noise process generates the $\mathbf{x}_i$ by randomly selecting a fraction of ϵ elements and replacing them with random values. Following this generative process, the negative log-likelihood is $R = \sum_i R_{i,\mathcal{L}_i}$, where the individual costs of assigning data item i to source set $\mathcal{L}_i$ are

$$R_{i,\mathcal{L}_i} \quad = \quad -\sum_{d=1}^{D} \log \Big((1-\epsilon)\,(1-v_{\mathcal{L}_i,d})^{x_{id}}\, v_{\mathcal{L}_i,d}^{1-x_{id}} + \epsilon\, r^{x_{id}}\,(1-r)^{1-x_{id}} \Big)\,. \tag{16}$$

Multi-Assignment Clustering (MAC) supports the simultaneous assignment of one data item to more than one source, i.e. the source sets can contain more than one element ($|\mathcal{L}_i| \geq 1$), while Single-Assignment Clustering (SAC) has the constraint $|\mathcal{L}_i| = 1$ for all i. Hence, when MAC has K sources and L different source combinations, the SAC model needs L independent sources for an equivalent model complexity. For MAC, $v_{\mathcal{L}_i,d} := \prod_{\lambda \in \mathcal{L}_i} v_{\lambda,d}$ is the product of all source parameters in assignment set $\mathcal{L}_i$, while for SAC, $v_{\mathcal{L}_i,d}$ is an independent parameter of the cluster indexed by $\mathcal{L}_i$. SAC thus has to estimate $L \cdot D$ parameters, while MAC uses the data more efficiently to only learn $K \cdot D$ parameters of the individual modes.

The model parameter ϵ is the mixture weight of the noise process, and r is the probability for a noisy bit to be 1. Fixing $\epsilon = 0$ corresponds to a generative model without noise process.

In summary, there are four model variants, each one defined by the constraints of its parameters: MAC models are characterized by $|\mathcal{L}_i| \geq 1,\ \mathbf{v} \in [0,1]^{K\cdot D}$, SAC models by $|\mathcal{L}_i| = 1,\ \mathbf{v} \in [0,1]^{L\cdot D}$; generative models without noise are described by $\epsilon = 0$ and its noisy version by $\epsilon \in [0,1[$.

5.2. Computation of the approximation capacity.

For the cost function in Eq. (16) the models. solutions in the hypothesis space. the hypothesis space is spanned by all possible assignments of objects to source combinations. A solution (a point in this hypothesis space) is encoded by the n source-sets $\mathcal{L}_i, i \in \{1,..,n\}$ with $|\mathcal{L}_i| \in \{1,..,K\}$. We explained in the last section that L has the same magnitude for all four model variants. Therefore, the hypothesis space of all four models equals in cardinality. In the following, we use the running index $\mathcal{L}$ to sum over all L possible assignment sets.

As the probabilistic model factorizes over the objects (and therefore the costs are a sum over object-wise costs $R(\boldsymbol{v}_{\mathcal{L}_i*}, \mathbf{x}_{i*}^{(q)})$ in Eq. (16)) we can conveniently sum over the entire hypothesis space by summing over all possible assignment sets for each object, similar as described in Section 3. The weight sums are then

$$\mathcal{Z}^{(q)} = \prod_{i=1}^{n} \sum_{\mathcal{L}=1}^{L} \exp\left(-\beta R(\boldsymbol{v}_{\mathcal{L}*}, \mathbf{x}_{i*}^{(q)})\right),\ q = 1,2\ , \tag{17}$$

$$\mathcal{Z}_{12} = \prod_{i=1}^{n} \sum_{\mathcal{L}=1}^{L} \exp\left(-\beta(R(\boldsymbol{v}_{\mathcal{L}*}, \mathbf{x}_{i*}^{(1)}) + R(\boldsymbol{v}_{\mathcal{L}*}, \mathbf{x}_{i*}^{(2)}))\right)\ . \tag{18}$$

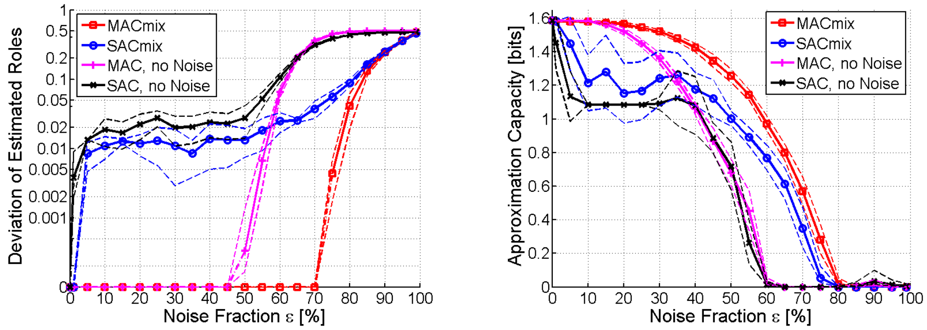
where the two datasets must be aligned before computing $R(\boldsymbol{v}_{\mathcal{L}*}, \mathbf{x}_{i*}^{(2)})$ such that $\mathbf{x}_{i*}^{(1)}$ and $\mathbf{x}_{i*}^{(2)}$ have a high probability to be generated by the same sources. In this particular experiment we guaranteed alignment by generation of the data. With real-world data one must use a mapping function as, for instance, in Frank et al. (2011).

The weight sums of the four model variants differ only in the combined source estimates $\boldsymbol{v}_{\mathcal{L}*}, \forall\mathcal{L}$. We train these estimates on the first dataset $\mathbf{x}^{(1)}$ prior to computing the mutual information Eq. (8). Having the formulas for the weight sums, one can readily evaluate the mutual information as a function of the inverse computational temperature β. We maximize this function numerically for each of the model variants.

5.3. Experiments

We investigate the dependency between the accuracy of the source parameter estimation and the approximation capacity. We choose a setting with 2 sources and we draw 100 samples from each source as well as from the combination of the two sources. The sources have 150 dimensions and a Hamming distance of 40 bits. To control the difficulty of the inference problem, the fraction ϵ of random bits varies between 0 and 0.99. The parameter of the Bernoulli-noise process is set to $r = 0.75$. The model parameters are then estimated by MAC and SAC both with and without a noise model. We use the true model order, i.e. $K = 2$ and $L = 3$ and infer the parameters by deterministic annealing Rose (1998).

The mismatch of the estimates to the true sources and the approximation capacity are displayed in Figures 3(a) and 3(b), both as a function of the noise fraction ϵ. Each method



(a) Source Estimation Mismatch

(b) Approximation Capacity

Figure 3: Error of source parameter estimation versus approximation capacity.

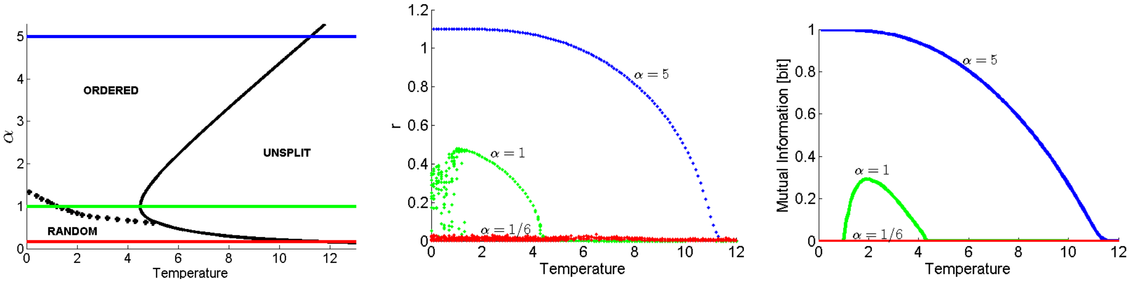
has very precise estimates up to a model-dependent critical noise level. For higher noise values, the accuracy breaks down. For both MAC and SAC, adding a noise model shifts this performance decay to a higher noise level. Moreover, MACmix estimates the source parameters more accurately than SACmix and shows its performance decay at a significantly increased noise levels. The approximation capacity (Fig. 3(b)) confirms this ranking. For noise-free data ($\epsilon = 0$), all four models attain the theoretical maximum of the approximation capacity, $\log_2(3)$ bits. As ϵ increases, the approximation capacity decreases for all models, but we observe vast differences in the sensitivity of the capacity to the noise level. Two effects decrease the capacity: inaccurately estimated parameters and (even with perfect estimates) the noise in the data that favors the assignment probabilities of an object to clusters to be more uniform (for $\epsilon = 1$ all clusters are equally probable). In conclusion, the wASC agrees with the ranking by parameter accuracy. We emphasize that parameter accuracy requires knowledge of the true source parameters while wASC requires only the data at hand.

6. Phase Transition in Inference

This section discusses phase transitions and learnability limits. We review theoretical results and show how the wASC principle can be employed to verify them.

6.1. Phase Transition of Learnability

Let the two centroids $\boldsymbol{\mu}_k$, $k = 1, 2$, be orthogonal to each other and let them have equal magnitudes: $|\boldsymbol{\mu}_1| = |\boldsymbol{\mu}_2|$. The normalized separation u is defined as $u := |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|/\sqrt{2\sigma_0}$, where σ_0 indicates the variance of the underlying Gaussian probability distributions with $\boldsymbol{\Sigma} = \sigma_0 \cdot \mathbf{I}$. We consider the asymptotic limit $D \to \infty$ while $\alpha := n/D$, σ_0 and u are kept finite as described in Barkai et al. (1993). In this setting, the complexity of the problem, measured by the Bayes error, is proportional to $\sqrt{1/D}$. Therefore, we decrease the distance between the centroids by a factor of $\sqrt{1/D}$ when going to higher dimensions in order to keep the problem complexity constant. Similar to the two dimensional study, we use annealed Gibbs sampling to estimate the centroids $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ at different temperatures. The theory of



(a) Phase Diagram (b) Overlap (c) Mutual Information

Figure 4: Experimental study of the overlap r and the mutual information $\mathcal{I}_\beta$ in different learnability limits. The problem complexity is kept constant while varying the number of objects per dimension α.

this problem is studied in Barkai and Sompolinsky (1994) and Witoelar and Biehl (2009). The study shows the presence of different phases depending on the values of stopping temperature and α. We introduce the same parameters as in Barkai and Sompolinsky (1994): The separation vector $\Delta\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)/2$, as well as the order parameters $s = \sigma_0|\Delta\hat{\boldsymbol{\mu}}|^2$ (the separation between the two estimated centers) and $r = \Delta\hat{\boldsymbol{\mu}} \cdot \Delta\boldsymbol{\mu}/u$ (the projection of the distance vector between the estimated centroids onto the distance vector between the true centroids). Computing these order parameters guides to construct the phase diagram. Thereby, we sample $n = 500$ data items from two Gaussian sources with orthogonal centroids $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and equal prior probabilities $\pi_1 = \pi_2 = 1/2$, and fix the variance σ_0 at 1/2. We vary α by changing the dimensionality D. To keep the Bayes error fixed, we simultaneously adapt the normalized distance. For different values of α we perform Gibbs sampling and infer the estimated centroids $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ at varying temperature. Then we compute the order parameters and thereby obtain the phase diagram shown in Fig. 4(a) which is consistent with the theoretical and numerical study in Barkai and Sompolinsky (1994):

Unsplit phase: $s = r = 0$. For high temperature and large α the estimated cluster centroids coincide, i.e. $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_2$.

Ordered split phase: $s, r \neq 0$. For values of $\alpha > \alpha_c = 4u^{-4}$, the single cluster obtained in the unsplit phase splits into two clusters such that the projection of the distance vector between the two estimated and the two true sources is nonzero.

Random split phase: $s \neq 0, r = 0$. For $\alpha < \alpha_c$, the direction of the split between the two estimated centers is random. Therefore, r vanishes in the asymptotic limits. The experiments also find such a meta-stability at low temperatures which correspond to the disordered spin-glass phase in statistical physics.

Therefore, as temperature decreases, different types of phase transitions can be observed:

1. $\alpha \gg \alpha_c$: Unsplit $\rightarrow$ Ordered. We investigate this scenario by choosing $D = 100$ and then $\alpha = 5$. The order parameter r in Fig. 4(b) shows the occurrence of such a phase transition.

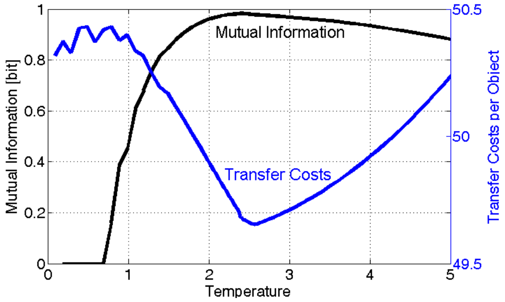


Figure 5: Expected transfer costs and approximation capacity when the number of observations per dimensions $\alpha = 5$. The Gibbs sampler is initialized with four centroids.

2. $\alpha \gtrapprox \alpha_c$: Unsplit → Ordered → Random. With $n = D = 500$, we then have $\alpha = 1$. The behavior of the parameter r is consistent with the phase sequence "Unsplit → Ordered → Random" as the temperature decreases. This result is consistent with the previous study in Barkai and Sompolinsky (1994).
3. $\alpha \ll \alpha_c$: Random phase. With the choice of $D = 3000, \alpha = 1/6$ then r is always zero. This means there is almost no overlap between the true and the estimated centroids.

As mentioned before, changing the dimensionality affects the complexity of the problem. Therefore, we adapt the distance between the true centroids to keep the Bayes error fixed. In the following, we study the approximation capacity for each of these phase transitions and compare them with the results we obtain in simulations.

6.2. Approximation Capacity of Phase Transition in Learnability Limits

Given the two datasets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ drawn from the same source, we calculate the mutual information between the first and the second datasets according to Eq. 13. We again numerically compute the mutual information $\mathcal{I}_\beta$ for the entire interval of β to obtain the approximation capacity (Eq. 9). Figure 4(c) shows this numerical analysis for the three different learnability limits. The approximation capacity reflects the difference between the three scenarios described above:

1. **Unsplit → Ordered**: The centroids are perfectly estimated. The approximation capacity attains the theoretical maximum of 1 bit at low temperature.
2. **Unsplit → Ordered → Random**: The strong meta-stability for low temperatures prevents communication. The mutual information is maximized at the lowest temperature above this random phase.
3. **Random**: The centroids are randomly split. Therefore, there is no information between the true and the estimated centroids over the entire temperature range. In this regime the mutual information is always 0 over all values of β.

We extend the study to a hypothesis class of 4 centroids, thus enabling the sampler to overfit. Using Gibbs sampling on $\mathbf{X}^{(1)} \in \mathbb{R}^{500\times100}$ (scenario (1)) under an annealing schedule, we compute the clustering $c(\mathbf{X}^{(1)})$. At each temperature, we then compute the mutual information and the transfer costs. In this way, we study the relationship between approximation capacity and generalization error in the asymptotic limits. Figure 5 illustrates the consistency of the costs of the transferred clustering solution with the approximation capacity. Furthermore, in the annealing procedure, the correct model order, i.e. $\hat{K} = 2$, is attained at the temperature that corresponds to the maximal approximation capacity.

7. Conclusion

Model selection and model order selection pose critical design issues in all unsupervised learning tasks. The principle of maximum approximation capacity (wASC) offers a theoretically well-founded approach to answer these questions. We have motivated this principle and derived the general form of the capacity. As an example, we have studied the approximation capacity of Gaussian mixture models (GMM). Thereby, we have demonstrated that the choice of the optimal number of Gaussians based on the approximation capacity coincides with the configurations yielding optimal generalization ability. *Weighted approximation set coding* finds the true number of Gaussians used to generate the data.

Weighted approximation set coding is a very general model selection principle which is applicable to a broad class of pattern recognition problems (for SVD see Frank and Buhmann (2011)). We have shown how to use wASC for model selection and model order selection in clustering. Future work will address the generalization of wASC to discrete continuous optimization problems, such as sparse regression, and to algorithms without cost functions.

Acknowledgments

This work has been partially supported by the DFG-SNF research cluster FOR916, by the FP7 EU project SIMBAD, and by the Zurich Information Security Center.

References

N. Barkai and H. Sompolinsky. Statistical mechanics of the maximum-likelihood density estimation. *Phys. Rev. E*, 50(3):1766–1769, 1994.

N. Barkai, H. S. Seung, and H. Sompolinsky. Scaling laws in learning of classification tasks. *Phys. Rev. Lett.*, 70(20):3167–3170, 1993.

Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In G. Lugosi and H.U. Simon, editors, *COLT'06, Pittsburgh, PA, USA*, pages 5–19, 2006.

Joachim M. Buhmann. Information theoretic model validation for clustering. In *International Symposium on Information Theory*, pages 1398 – 1402. IEEE, 2010.

Joachim M. Buhmann. Context sensitive information: Model validation by information theory. In *MCPR 2011*, volume 6718 of *LNCS*, pages 21–21. Springer, 2011.

Kenneth P. Burnham and David R. Anderson. *Model selection and inference: a practical information-theoretic approach, 2nd ed.* Springer, New York, 2002.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley, 2006.

Sandrine Dudoit and Jane Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7), 2002.

Mario Frank and Joachim M. Buhmann. Selecting the rank of SVD by maximum approximation capacity. In *ISIT 2011.* IEEE, 2011.

Mario Frank, Morteza Chehreghani, and Joachim M. Buhmann. The minimum transfer cost principle for model-order selection. In *ECML PKDD '11: Machine Learning and Knowledge Discovery in Databases*, volume 6911 of *Lecture Notes in Computer Science*, pages 423–438. Springer Berlin / Heidelberg, 2011.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE PAMI*, 6(6):721–741, 1984.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Verlag, New York, 2008.

Tilman Lange, Mikio Braun, Volker Roth, and Joachim M. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16(6):1299–1323, June 2004.

Kenneth Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *IEEE PAMI*, 86(11):2210–2239, 1998.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

Susanne Still and William Bialek. How many clusters? an information-theoretic perspective. *Neural Computation*, 16:24832506, 2004.

Andreas P. Streich, Mario Frank, David Basin, and Joachim M. Buhmann. Multi-assignment clustering for Boolean data. In *ICML'09*, pages 969–976, 2009.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

Vladimir N. Vapnik. *Statistical learning theory.* Wiley, New York, 1998.

Aree Witoelar and Michael Biehl. Phase transitions in vector quantization and neural gas. *Neurocomputing*, 72(7-9):1390–1397, 2009.